



US006055540A

**United States Patent** [19][11] **Patent Number:** 6,055,540

Snow et al.

[45] **Date of Patent:** Apr. 25, 2000

[54] **METHOD AND APPARATUS FOR CREATING A CATEGORY HIERARCHY FOR CLASSIFICATION OF DOCUMENTS**

[75] **Inventors:** William A. Snow, Redwood City;  
Joseph D. Mocker, Cupertino, both of Calif.

[73] **Assignee:** Sun Microsystems, Inc.

[21] **Appl. No.:** 08/874,567

[22] **Filed:** Jun. 13, 1997

[51] **Int. Cl.<sup>7</sup>** ..... G06F 17/30

[52] **U.S. Cl.** ..... 707/103; 707/100; 707/3

[58] **Field of Search** ..... 707/100, 102,  
707/103, 104, 3

5,706,496	1/1998	Noguchi et al. ....	395/603
5,721,910	2/1998	Unger et al. ....	707/100
5,768,578	6/1998	Kirk et al. ....	395/611
5,778,362	7/1998	Deerwester ....	707/5
5,778,372	7/1998	Cordell et al. ....	707/100
5,781,914	7/1998	Stork et al. ....	707/506
5,787,425	7/1998	Bigus ....	707/6
5,802,518	9/1998	Karaev et al. ....	707/9
5,806,068	9/1998	Shaw et al. ....	707/103
5,809,340	9/1998	Bertone et al. ....	395/878
5,812,995	9/1998	Sasaki et al. ....	707/1
5,813,014	9/1998	Gustman ....	707/103
5,835,712	11/1998	DuFresne ....	709/203
5,838,965	11/1998	Kavanaugh et al. ....	707/103
5,862,325	1/1999	Reed et al. ....	709/201

**Primary Examiner**—Thomas G. Black

**Assistant Examiner**—Charles L. Rones

**Attorney, Agent, or Firm**—D'Alessandro & Ritchie

[57] **ABSTRACT**

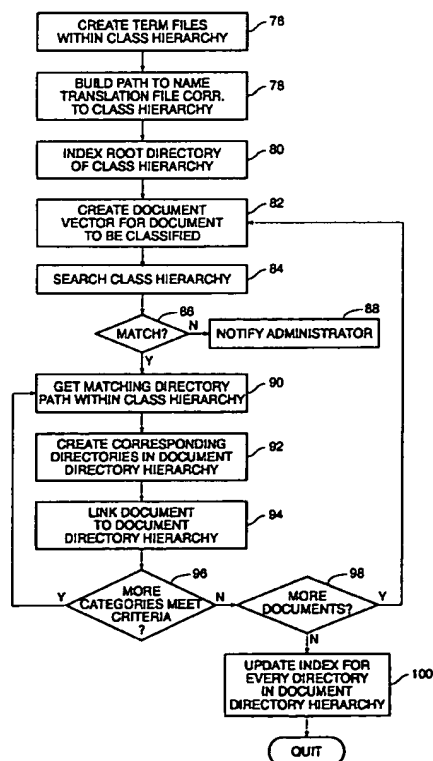
A method for creating a class hierarchy containing categories for classification of documents. The class hierarchy is initialized to contain a root category node within a tree data structure. The root category node is defined by a user-defined category name. The class hierarchy is displayed to assist a user in entering a command for manipulating the class hierarchy. A user may select a category command, resulting in the class hierarchy containing a plurality of category nodes. In addition, a user may select a terms command to manipulate terms defining one of the plurality of category nodes.

24 Claims, 6 Drawing Sheets

[56] **References Cited**

# U.S. PATENT DOCUMENTS

5,162,992	11/1992	Williams .....	364/419
5,297,249	3/1994	Bernstein et al. ....	345/356
5,301,319	4/1994	Thrumman et al. ....	707/103
5,317,646	5/1994	Sang et al. ....	382/175
5,333,237	7/1994	Stefanopoulos et al. ....	706/11
5,355,472	10/1994	Lewis .....	707/101
5,418,946	5/1995	Mori .....	395/600
5,442,778	8/1995	Pedersen et al. ....	395/600
5,463,773	10/1995	Sakakibara et al. ....	707/102
5,568,640	10/1996	Nishiyama et al. ....	395/600
5,634,051	5/1997	Thomson .....	395/605
5,649,186	7/1997	Ferguson .....	395/610



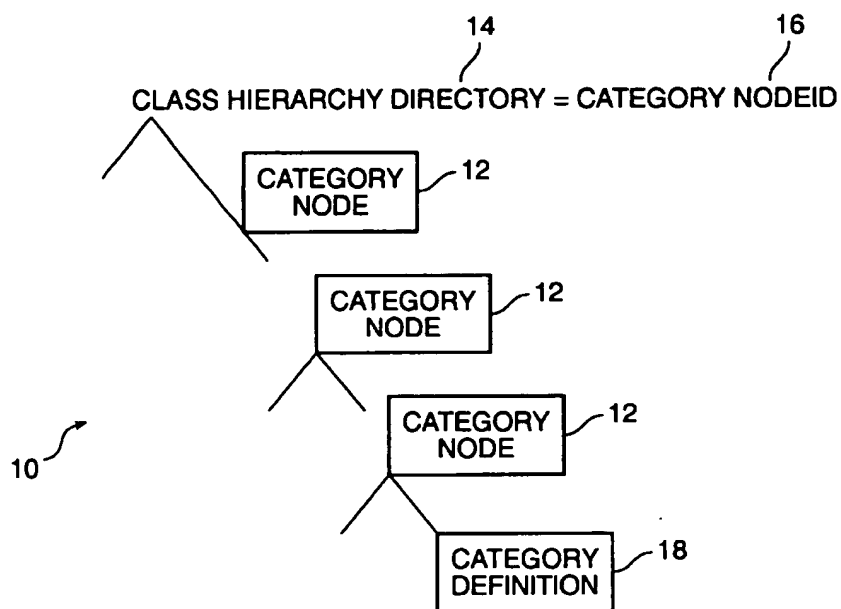


FIG. 1

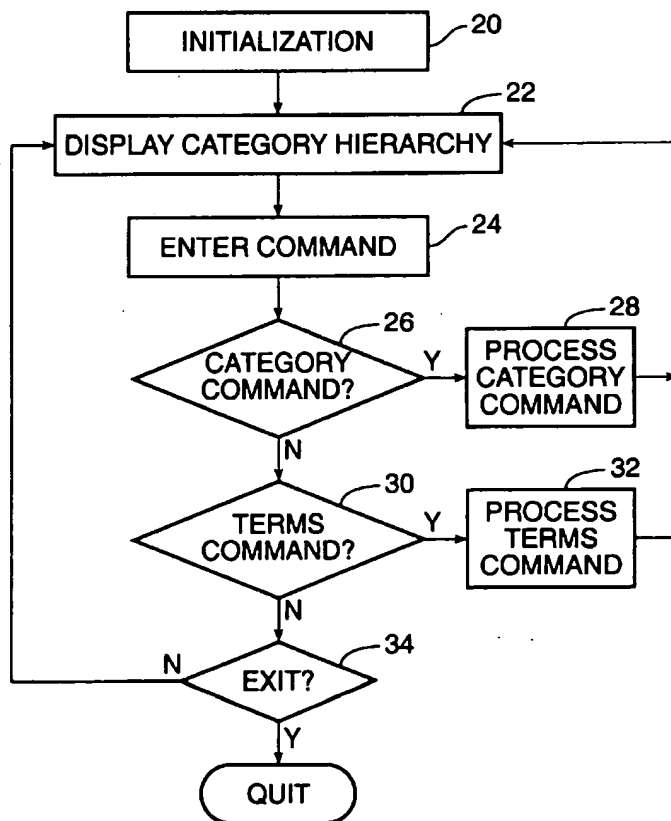


FIG. 2

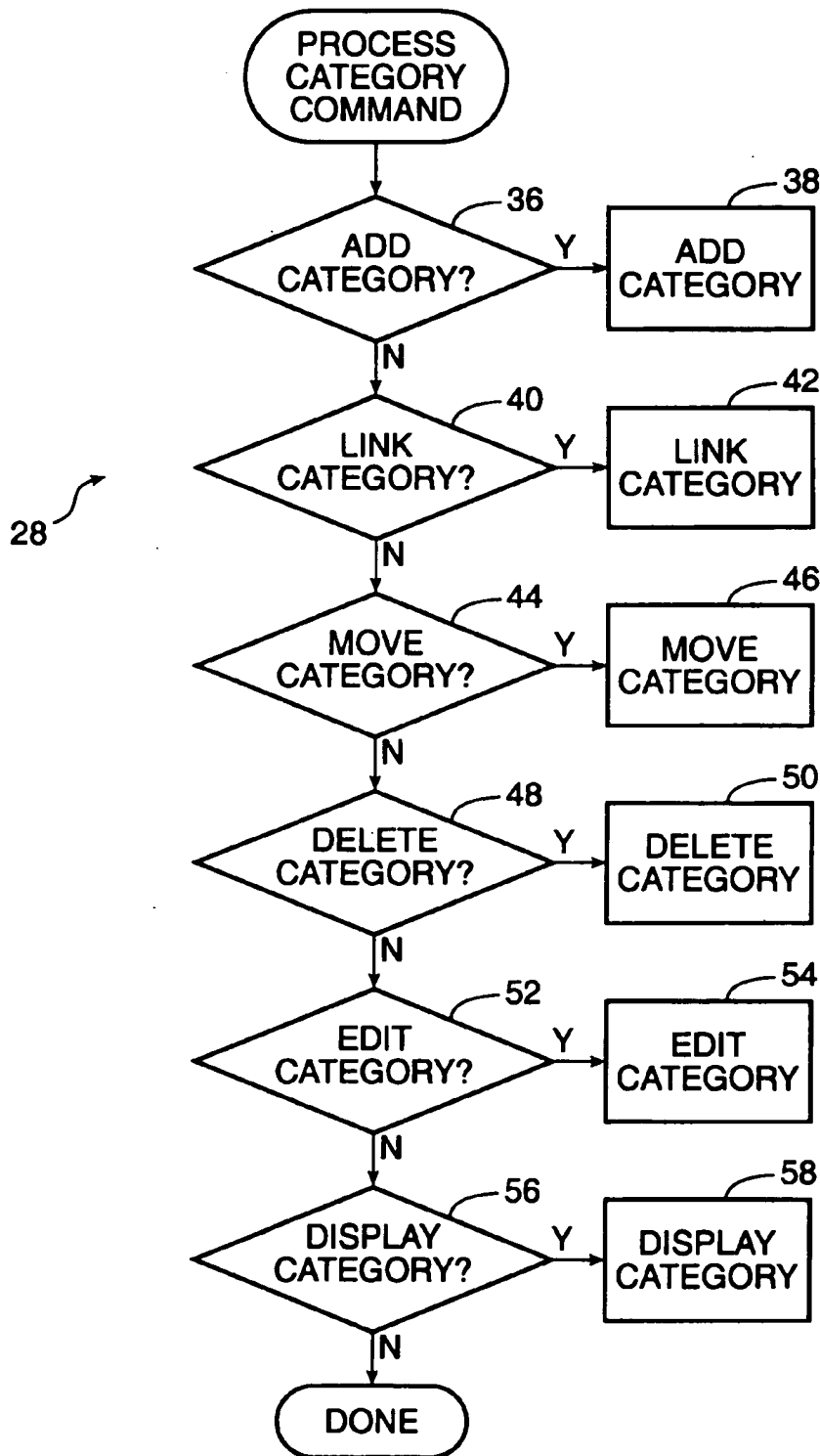


FIG. 3

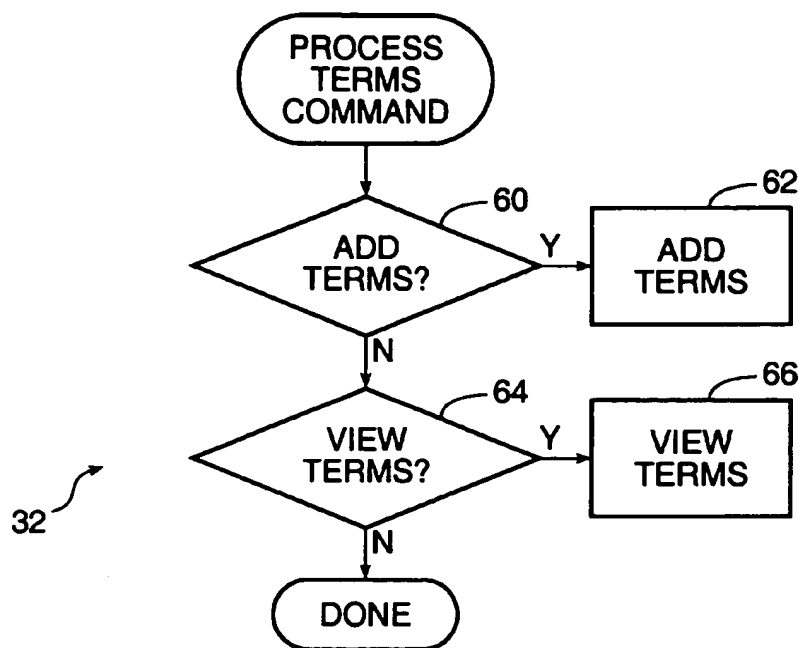


FIG. 4

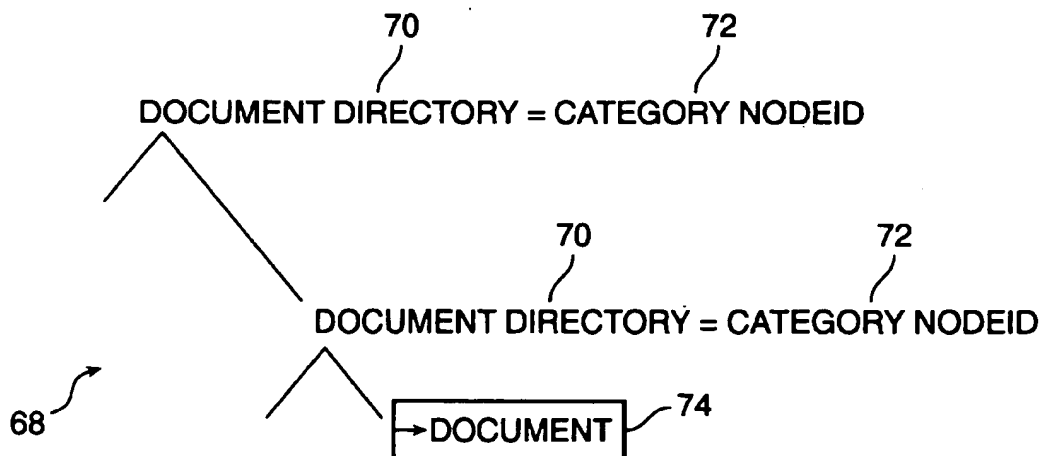


FIG. 5

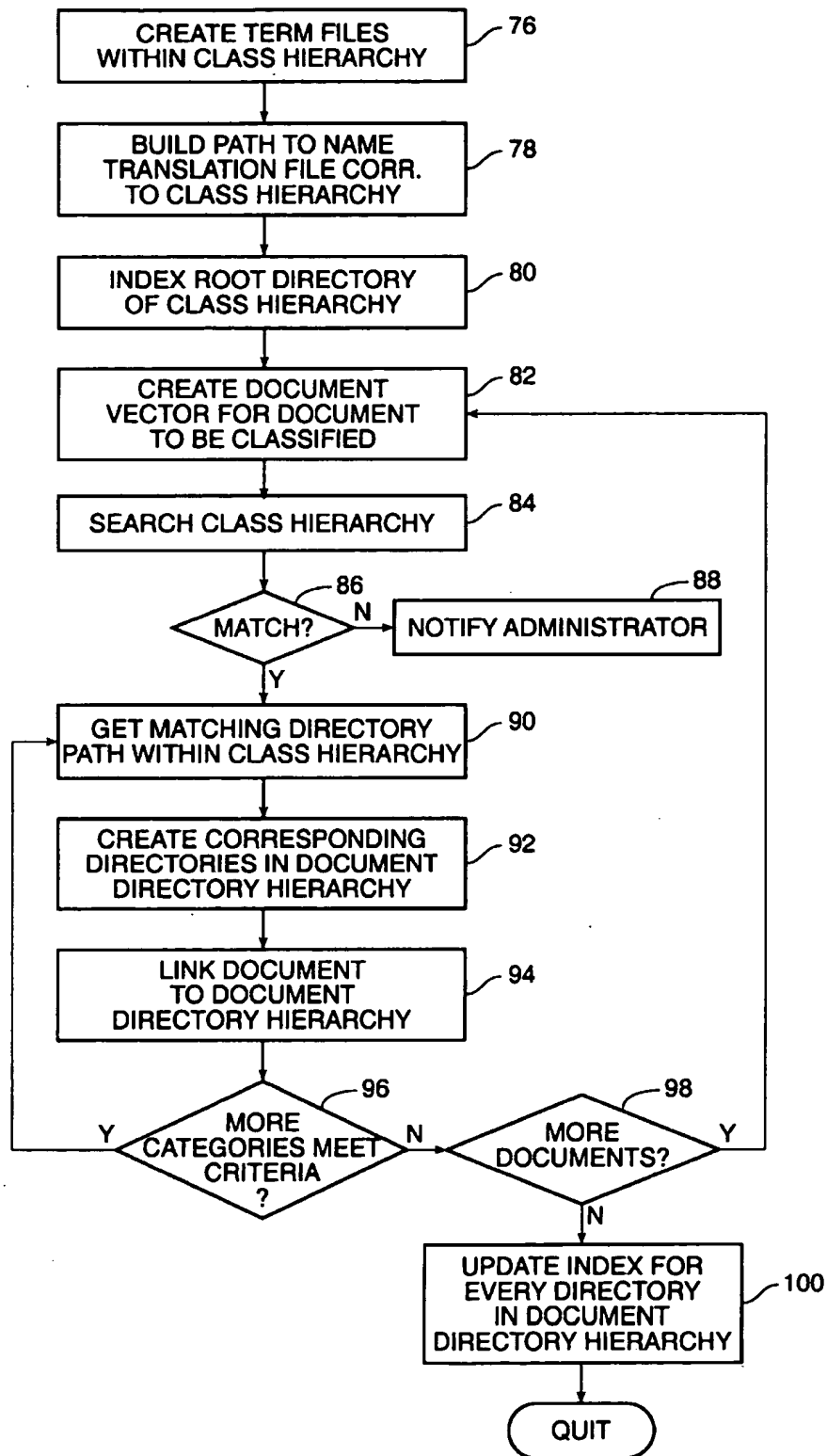


FIG. 6

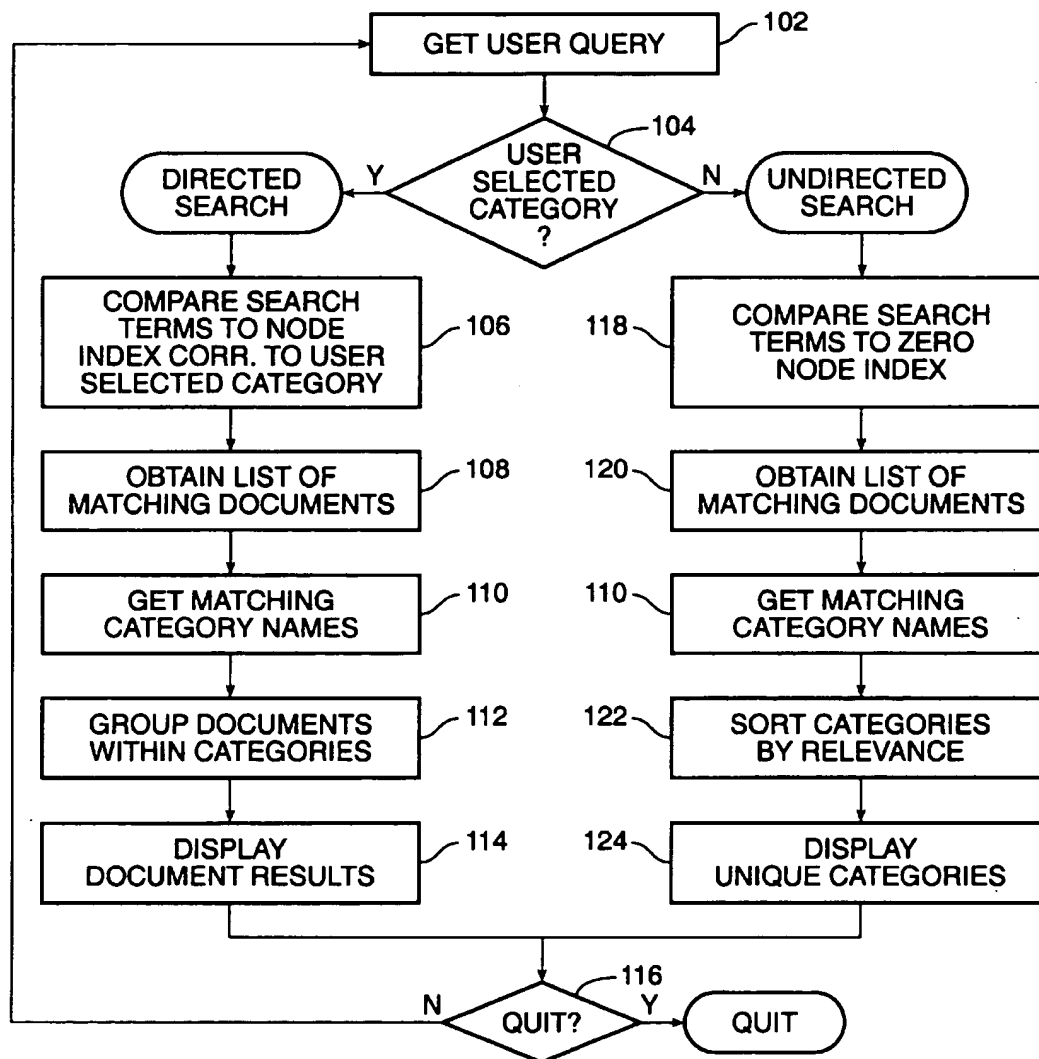


FIG. 7

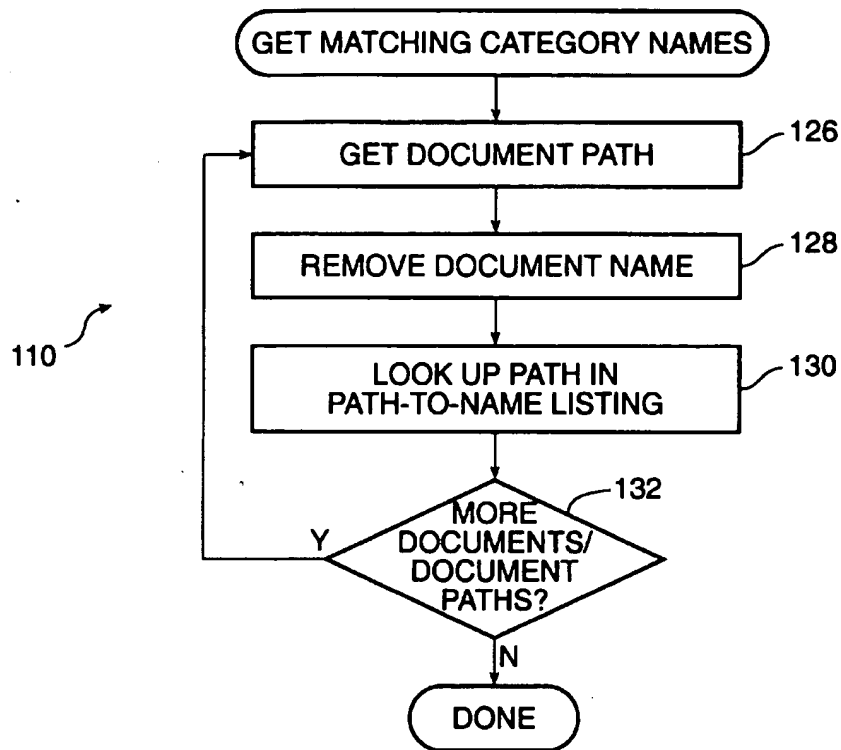


FIG. 8

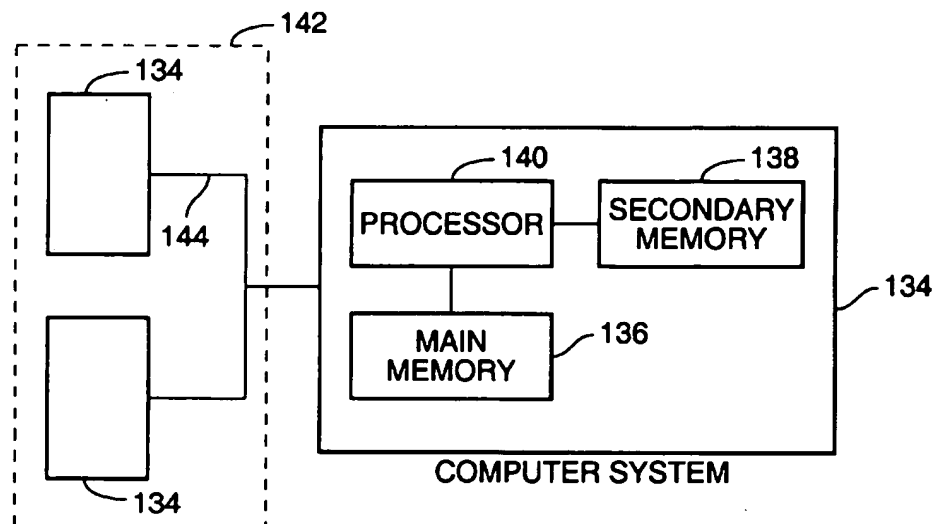


FIG. 9

# METHOD AND APPARATUS FOR CREATING A CATEGORY HIERARCHY FOR CLASSIFICATION OF DOCUMENTS

## BACKGROUND OF THE INVENTION

### 1. Field of the Invention

The embodiments of the present invention relate to a method and apparatus for creating a class hierarchy for classification of documents. More particularly, the embodiments of the present invention relate to a method and apparatus for creating document categories within a class hierarchy and creating category definitions defining the document categories to allow classification of documents within the document categories.

### 2. Background of the Invention

Accurate classification of information is typically accomplished through the use of various classes and corresponding criteria for classification. Moreover, various types of hierarchies may be used for information storage. For example, systems using tree data structures are commonly used to store related information.

Retrieval of documents from such a system can be performed most efficiently when the documents are properly classified. A need exists in the prior art for a method and apparatus for creating a class hierarchy containing classes and criteria for classification. Implementing such a class hierarchy in a computerized document classification system would provide for efficient and accurate categorization of documents. Moreover, it would be extremely beneficial if such a system were made available to multiple users via a communications network such as a computer network.

## BRIEF DESCRIPTION OF THE INVENTION

A method and apparatus for creating a class hierarchy containing categories for classification of documents is provided. The class hierarchy is then used with a document classifier to allow classification of documents within the categories. In accordance with the invention, a plurality of category nodes and corresponding terms defining each of the plurality of category nodes are stored within a class hierarchy. The class hierarchy is initialized to contain a root category node within a tree data structure, and the root category node is defined by a user-defined category name. The class hierarchy is displayed to assist a user in entering a command for manipulating the class hierarchy. First, a user may select a category command, resulting in the class hierarchy containing a plurality of category nodes. The category command may comprise one of several commands. These commands include the capability to add a category node, link a first category node to a second category node, move a category node, delete a category node, edit a category node, or display information defining a category node. Second, a user may select a terms command to manipulate terms defining one of the plurality of category nodes. Therefore, terms can be added to one of the plurality of category nodes or viewed according to user-selected criteria.

## BRIEF DESCRIPTION OF THE DRAWINGS

The following figures are illustrative and not intended to limit the scope of the invention.

FIG. 1 illustrates a class hierarchy according to one embodiment of the present invention.

FIG. 2 is a flow diagram of the main program loop utilized in creation of the class hierarchy according to an embodiment of the present invention.

FIG. 3 is a flow diagram of the process category command procedure shown in the main program loop of FIG. 2 according to one embodiment of the present invention.

FIG. 4 is a flow diagram of the process terms command procedure shown in the main program loop of FIG. 2 according to one embodiment of the present invention.

FIG. 5 illustrates a document directory hierarchy according to one embodiment of the present invention.

FIG. 6 is a flow diagram of the main procedure utilized in creation of the document directory hierarchy according to one embodiment of the present invention.

FIG. 7 is a flow diagram illustrating a method for searching the document directory hierarchy according to one embodiment of the present invention.

FIG. 8 illustrates a method for retrieving matching category names corresponding to directory paths shown in FIG. 7 according to one embodiment of the present invention.

FIG. 9 illustrates a block diagram of an embodiment of a computer system implementing the present invention.

## DETAILED DESCRIPTION OF THE EMBODIMENTS OF THE INVENTION

Those of ordinary skill in the art will realize that the following description of the embodiments of the present invention is illustrative only and is not intended to be in any way limiting. Other embodiments of the invention will readily suggest themselves to such skilled persons from an examination of the within disclosure.

The embodiments of the present invention provide for automatic document classification within user-defined categories. A user can then interactively search for documents according to search terms within the user-defined categories. Documents are ranked according to relevance, and a user specified number of documents which are most relevant are returned. According to one embodiment, the present invention is made available to multiple users via a network.

While the embodiments of the present invention are of broad applicability and can be used in a variety of contexts, an embodiment of the present invention is designed specifically to interface with Fulcrum™, an information retrieval system designed for use by application developers. Fulcrum™ is available from Fulcrum Technologies™, Inc., located at 785 Carling Ave, Ottawa, Canada K1S5H4, (613) 238-1761. The functions provided by Fulcrum™ include text search, indexing and retrieval capabilities. Indexing creates a summary of all documents within a desired directory and subdirectories. When a directory is indexed, Fulcrum™ creates an index comprising a document vector for each document within the selected directory. To create each document vector, each document is split into terms and a weight is associated with each of the terms. The term weights are based upon frequency of occurrence of each term within the document. The weights, therefore, are higher when a term exists multiple instances within the document.

Method for Creating a Class Hierarchy

Referring now to FIG. 1, a class hierarchy 10 according to one embodiment of the present invention is shown. The class hierarchy 10 comprises at least one level of categories. These categories may further comprise sub-categories. The categories are stored in category nodes 12 within a tree data structure. According to one embodiment of the present invention, each of the category nodes 12 corresponds to a class hierarchy directory 14 equivalent to a category NodeID 16. Each category not further comprising sub-categories will herein be referred to as a leaf node, or leaf category. Each



leaf category comprises a category definition 18 defining the leaf category. According to one embodiment of the present invention, the category definition 18 comprises two groups of data. The first group of data contains descriptive terms defining the corresponding leaf category. The second group of data contains portions of documents which have been classified by a user as being relevant to the leaf category. Each category further comprising subcategories is defined by the terms corresponding to all subcategories within the category. The class hierarchy 10 is stored in a class hierarchy database.

According to one embodiment of the present invention, the category node 12 includes the following fields:

**Category name**—The category name is used for display purposes.

**Node type**—There are three possible node types: "normal", "see" and "see also". "See also" indicates that a category node can be added at this level, but lists alternate locations in which it may also be placed. "See" indicates that a category node cannot be added at this level, and gives alternate categories in which to place the new category. "Normal" indicates that the node is either a branch or a leaf node. The node type is set at time of creation of the node.

**NodeID**—When a node is created, an integer NodeID is associated with the node. According to one embodiment of the present invention, each directory within the class hierarchy comprises a directory name equivalent to the NodeID of the corresponding category. The NodeID is used rather than the category name, since each category name may contain characters that would result in an invalid directory name.

**ParentID**—The ParentID field contains an integer indicating the NodeID of the parent node.

**LinkID**—The LinkID field is used for "see" and "see also" node types. The LinkID contains an integer indicating a NodeID of a desired reference node.

**Entered by**—The entered by field comprises a user ID entered by the user. This field provides a means for tracking system updates and creation of system errors.

One of ordinary skill in the art will recognize that the category node may comprise fewer or additional fields.

Referring now to FIG. 2, a flow diagram of the main program loop utilized in creation of the class hierarchy according to one embodiment of the present invention is shown. At step 20, initialization of the class hierarchy is performed. Initially, the class hierarchy comprises a root category, or root node. In addition, an initial set of descriptive terms, including document portions, are provided to a user. The set of initial terms and document portions are generated from various sources (i.e., keyword tables, search log files).

The class hierarchy is displayed at step 22. To display the updated class hierarchy, the class hierarchy is traversed. First, the root is found by finding the node having no parent. The NodeID of the root is then obtained. Second, it is determined whose parent is the NodeID of the root. The second step is iteratively performed with the NodeID of the current node until the LinkID indicates that the node contains no children. Since categories and terms cannot be added to a node unless a node contains only "see also" subnodes, this is indicated in the display. For example, if a node contains only "see also" subnodes, this node will be displayed to the user. However, if a node contains a subnode having a "normal" or "see" node type, the node will not be displayed or this limitation will otherwise be indicated to the user. Alternatively, the class hierarchy may be displayed at a later step.

Several category commands are available to create the class hierarchy. These commands include the capability to add a category node, link a first category node to a second category node, move a category node, delete a category node, edit a category node, or display information defining a category node. Manipulating nodes in a tree data structure is known in the art of software development.

At step 24, a user-selected command is entered. If the user-selected command is determined to be a category command at step 26, the appropriate category command is processed at step 28, and the routine returns to step 22. If it is determined that the user-selected command is not a category command at step 26, it is next determined whether the user-selected command is a terms command at step 30. If the user-selected command is a terms command, the appropriate terms command is processed at step 32, and the loop is repeated at step 22. The loop is repeated at step 22 until the user chooses to exit the main program loop at step 34. Alternatively, the steps 26–32 may be ordered in various ways to achieve the same result.

Referring now to FIG. 3, a flow diagram of the process category command procedure 28 shown in FIG. 2 according to one embodiment of the present invention is illustrated. If the user-selected command is determined to be an add category command at step 36, the add category command is performed at step 38. This command is used to add a new category by name, adding a child node to an existing category. Add is not available if the node is a leaf node or the node contains a "see" link. Add is available if a node contains only "see also" sub-nodes. The first category, or node, comprises a root directory, or 0 node. When a new category is added, a new category node is created containing a user-defined category name, a node type, a unique NodeID corresponding to the user-defined category name, a ParentID corresponding to a NodeID of a parent category node, a LinkID, and a UserID. This category node is then stored within the parent category node.

If the user-selected command is not an add category command, at step 40 it is determined if the user-selected command is a link category command. If the user-selected command is a link category command, the link category command is performed at step 42. The link category command allows a category to refer to another category using "see" or "see also". "See also" indicates that one or more other categories contain related information. "See" indicates that this category name cannot contain sub-categories. Link is not available at the root node. Furthermore, the link command is not available if a node already contains a "see" link.

If the user-selected command is not a link-category command, at step 44 it is determined if the user-selected command is a move category command. If the user-selected command is a move category command, the move category command is performed at step 46. The move category command allows a category to be moved to another location within the class hierarchy.

If the user-selected command is not a move-category command, at step 48 it is determined if the user-selected command is a delete category command. If the user-selected command is a delete category command, the delete category command is performed at step 50. The delete category command deletes a category and any subtree. This command is not available at the root level.

If the user-selected command is not a delete category command, at step 52 it is determined if the user-selected command is an edit category command. If the user-selected command is determined to be an edit category command, the

edit category command is performed at step 54. The edit category command allows a category to be renamed.

If the user-selected command is not an edit category command, at step 56 it is determined if the user-selected command is a display category command. If the user-selected command is determined to be a display category command, the display category command is performed at step 58. This command displays node information corresponding to a particular category. According to one embodiment of the present invention, ParentID and LinkID information are not displayed. If the user-selected command is not a display category command, the process category command routine is completed and the program returns to the main loop of FIG. 2. One of ordinary skill in the art will recognize that steps 36-58 may be performed in an alternate order.

According to one embodiment of the present invention, the category definition for each leaf category in the class hierarchy is stored in a terms database. The terms database comprises descriptive terms and portions of documents defining all leaf categories. Each of the descriptive terms include a reference corresponding to at least one leaf category. Similarly, each of the portions of documents include a reference corresponding to at least one leaf category, document name, type of document, fields, or portions, included within the document (i.e., synopsis), and indexing information indicating which fields, or portions, are to be extracted during indexing. Therefore, a category definition may be defined by different portions of documents depending upon the type of each document.

Referring now to FIG. 4, a flow diagram of the process terms command procedure 32 shown in FIG. 2 according to one embodiment of the present invention is presented. If the user-selected command is determined at step 60 to be an add terms command, the add terms command is performed at step 62. The "add terms" command allows terms or portions of documents to be added to the category definition of a particular category. According to one embodiment of the present invention, the category must comprise a leaf category. A user can add a term from the initial set of terms, or add additional user-defined terms or document portions to an appropriate category. Multiple terms separated by commas can be entered. A confirmation dialog is presented before a term is added or deleted. Terms of a branch node may be edited if the only type of sub-nodes are "see also" links. This command provides a user with the ability to weight the descriptive terms which define a category. Terms, as well as each hand-classified document portion, can be given different weights. This is performed by creating multiple copies of a selected term or hand-classified document portion within the terms database.

If the user-selected command is not an add terms command, it is next determined if the user-selected command is a view terms command at step 64. If the user-selected command is a view terms command, the view terms command is performed at step 66. The "view terms" command allows a user to view the category definition of a particular category. This command is available only at the root node. This command is used to display all terms according to selection criteria. For example, a user may display the terms entered by anyone or display the terms entered by a user. If the user-selected command is not a view terms command, the process terms command procedure is complete and the program returns to the main loop of FIG. 2. One of ordinary skill in the art will recognize that steps 60-66 may be performed in an alternate order. Method for Classifying a Document within the Class Hierarchy

Once the class hierarchy has been created, documents may be classified within the class hierarchy. According to the embodiments of the present invention, a document is classified based on content within the categories within the class hierarchy. All documents entered into the system are classified within one or more categories. The disclosed method for classification results in correct placement of documents by the classifier approximately 80% of the time.

Referring now to FIG. 5, a document directory hierarchy 68 according to one embodiment of the present invention is shown. The document directory hierarchy 68 comprises a plurality of document directories 70. Each of the directories 70 corresponds to a category within the class hierarchy. According to one embodiment of the present invention, each of the document directories 70 is equivalent to a category NodeID 72 of the corresponding category. Once a document 74 is classified within the class hierarchy, the document 74 is stored within the document directory 68 as shown.

Referring now to FIG. 6, a flow diagram of the main procedure utilized in creation of the document directory hierarchy according to one embodiment of the present invention is shown. Initially, at step 76, a terms file is created for each category within the class hierarchy. A category definition for each leaf node is extracted from the terms database and stored in a terms file within the category directory in the class hierarchy. Therefore, each terms file will contain all terms and portions of documents defining the particular category.

Next, at step 78, a path-to-name listing containing a directory path to category name translation for each directory within the class hierarchy is created and stored in a translation file. The categories are extracted from the class hierarchy database by traversing the class hierarchy from the root, finding all the children successively until a leaf node is hit. This process is repeated for all children. Each directory within the directory hierarchy comprises a directory name equivalent to the NodeID of the corresponding category.

Next, at step 80, indexing via Fulcrum™ is performed. Since each leaf node initially contains only one term document, or term file, indexing is performed on the class hierarchy which contains the terms files. Fulcrum extracts the terms from each term file and weights each term according to frequency of occurrence. Traversing subdirectories is standard for the Fulcrum™ search engine. Fulcrum™ is used to index the 0 node, or root directory, which results in indexing of each subdirectory within the 0 node, creating a zero node index. Then, for each term within all term files, Fulcrum™ creates a term vector for each of the most common terms of the document, and corresponding positioning information. Indexing creates an index file which contains all term file document vectors.

Next, a document vector is created at step 82 for a document to be classified. The document vector is created via Fulcrum to allow classification through comparison with document vectors created in step 80.

Next, at step 84, a text search is performed within the class hierarchy to classify a document. The document to be classified is compared against all leaf node data to determine the appropriate category placement for the document. This is performed by comparing document vectors. The document vector created in step 82 is compared via Fulcrum™ to the term file document vectors created during the indexing step 80 to determine appropriate category placement. As a result, Fulcrum™ returns a relevance ranking. A top percentage of the rankings are utilized to determine appropriate categories for the document. This relevance percentage is configurable within the system. The document matches one or more

categories if it meets the user-defined criteria, or configured relevance percentage. A result of the search is a list of matching category names.

At step 86, if the document does not match any of the categories, the system administrator is notified at step 88. If, however, the document meets the user-defined criteria, a matching directory path within the class hierarchy is obtained at step 90 for one of the matching category names utilizing the path-to-name translation file.

Next, any necessary directories corresponding to the matching directory path are created within the document directory hierarchy at step 92. Therefore, the directory structure within the class hierarchy will not necessarily be equivalent to that of the document directory hierarchy. As a result, only directories containing documents will be created within the document directory hierarchy. Alternatively, a directory structure equivalent to that of the class hierarchy may be created in a prior step.

Next, at step 94, the document is added to a leaf directory within the document directory hierarchy corresponding to the matching directory path. According to one embodiment of the present invention, the document is symbolically linked via Unix to the directory corresponding to the matching category.

At step 96, it is determined whether there are more matching directory paths to which the document must be linked. If there are more matching directory paths, a matching directory path for the next matching category is retrieved at step 90, and the loop is repeated.

Once the document has been linked within the document directory hierarchy, more documents may be classified by a user at step 98. When no more categories meet the user-defined criteria, more documents may be classified by a user. When all documents have been classified, each directory within the document directory hierarchy is indexed at step 100. When a branch node is indexed, all documents in sub-nodes of that node are indexed. Thus, only indexes corresponding to modified directories and the parent nodes of the modified directories will be updated. Alternatively, only modified directories and any parent nodes of the modified directories may be indexed. If there are more documents to classify at step 98, a document vector is created at step 82 and the loop is repeated.

Method for Searching the Document Directory Hierarchy

Referring now to FIG. 7, a flow diagram illustrating a method for searching the document directory hierarchy according to one embodiment of the present invention is shown. According to one embodiment of the present invention, two search methods are provided for searching the document directory hierarchy in response to a user query. According to a first method, an undirected search, a user query may comprise one or more search terms. In addition, after search results are obtained, the user can modify the original search terms to further limit the search. According to a second method, a directed search, a user query may comprise one or more search terms and a selected category to provide a more limited search. Once results are obtained, the user can then select one or more categories or modify the search terms to run a more limited search. According to one embodiment of the present invention, the user specifies a number of documents desired.

At step 102, a user query is obtained. The user query comprises a number of documents desired and one or more search terms. In addition, the user query may include a user selected category.

Next, at step 104, if the user query includes a user selected category, a directed search is performed. At step 106, the

search terms are compared to each of the relevant document vectors created by the document indexing of the document directory hierarchy. Since the search is directed, the relevant document vectors are the document vectors within the index corresponding to the desired category.

At step 108, a list of document directory paths is obtained through Fulcrum™. Each of the document directory paths includes a matching document name and a directory path corresponding to a matching document. The relevant documents are ranked according to relevance using a statistical ranking provided by Fulcrum™. According to one embodiment of the present invention, the user-specified number of document directory paths which are most relevant are selected. Then, matching category names are obtained at step 110.

According to the directed search, the documents are grouped within the matching category names at step 112. Next, information corresponding to each document is displayed by category at step 114. According to one embodiment of the present invention, the document information includes a synopsis and document link. Upon completion of the directed search, a user may choose to quit at step 116. If the user does not choose to quit, the user enters another user query at step 102 and the loop is repeated.

If at step 104, if the user query does not include a user selected category, an undirected search is performed. At step 118, the search terms are compared to each of the relevant document vectors created by the document indexing of the document directory hierarchy. Since the search is undirected, the relevant document vectors are the document vectors within the zero node index.

At step 120, a list of document directory paths is obtained through Fulcrum. Each of the document directory paths includes a matching document name and a directory path corresponding to a matching document. The relevant documents are ranked according to relevance using a statistical ranking provided by Fulcrum™. Then, matching category names are obtained at step 110.

According to the undirected search, all relevant category names obtained in step 110 are sorted by relevance at step 122. Duplicate category names are removed from each of the sorted relevant category names, and the unique sorted relevant category names are displayed at step 124. Upon completion of the undirected search, a user may choose to quit at step 116. If the user does not choose to quit, the user enters another user query at step 102 and the loop is repeated. Therefore, the user may select appropriate categories, alter the search terms, and re-run the search.

Referring now to FIG. 8, the method for retrieving matching category names 110 corresponding to directory paths shown in FIG. 7 according to one embodiment of the present invention is presented. At step 126, a matching document directory path is obtained from the matching document directory paths. Next, at step 128, a document name is removed from the document directory path to obtain a directory path. Next, at step 130, a corresponding category name is obtained by performing a search, or look up, for the directory path in the path-to-name listing to obtain a category name. At step 132, if there are more matching document paths, the loop is repeated at step 126. However, if none of the matching document paths remain, each of the matching category names have been retrieved.

Referring now to FIG. 9, a block diagram of an embodiment of a computer system 134 implementing the present invention is shown. According to this embodiment, the present invention is stored in a main memory 136 or a secondary memory 138 of the computer system 134 for use

by a processor 140. The computer system 134 may be connected to a computer network 142 through transmission lines 144. Those of ordinary skill in the art will readily recognize that the present invention may also be used in a standalone computer system, which is by definition not part of a computer network.

#### Alternative Embodiments

According to an alternative embodiment, the search may be performed on the index created from the terms files within the class hierarchy rather than the index created from the documents in the document directory hierarchy. However, although this method is more efficient than other embodiments, it allows a category which contains no documents to be displayed to the user. Since the user may select a category which contains no documents, this may be confusing to the user.

According to another alternative embodiment, the search may be performed on the index created from the terms files within the class hierarchy in addition to the index created from the documents. This would be helpful to find related categories of information.

Although illustrative embodiments and applications of this invention are shown and described herein, many variations and modifications are possible which remain within the concept, scope, and spirit of the invention, and these variations would become clear to those of skill in the art after perusal of this application. The invention, therefore, is not to be limited except in the spirit of the appended claims in light of their full scope of equivalents.

What is claimed is:

1. A method for creating a class hierarchy for categorization of documents within a memory, the class hierarchy for use with a document classification system capable of classifying a document based on content within the class hierarchy, the method comprising:

initializing the class hierarchy, the class hierarchy having a root category node within a tree data structure, the root category node having a user-defined category name;

displaying the class hierarchy;

accepting a user-selected command for manipulating the class hierarchy;

processing a category command in response to the user-selected command having a first predefined state, causing the class hierarchy to contain a plurality of category nodes, said processing the category command further comprising:

storing a category name in one of the plurality of category nodes, wherein each of the plurality of category nodes corresponds to a unique directory;

storing a NodeID within one of the plurality of category nodes, the NodeID defining the unique directory;

storing a nodetype within one of the plurality of category nodes, the nodetype when having a predefined type allowing a new category node to be added to a selected one of the plurality of category nodes, and otherwise preventing the new category node from being added to the selected one of the plurality of category nodes;

storing a ParentID within one of the plurality of category nodes, the ParentID indicating a NodeID of a parent category node; and

storing a LinkID within a first one of the plurality of category nodes, the LinkID indicating a NodeID of a second one of the plurality of category nodes when the nodetype is of a predefined type; and

processing a terms command in response to the user-selected command having a second predefined state,

the terms command manipulating terms defining one of the plurality of category nodes.

2. The method according to claim 1, said displaying further comprising displaying information corresponding to at least one of the plurality of category nodes within the class hierarchy, the information indicating a nodetype for the at least one of the plurality of category nodes.

3. The method according to claim 1, wherein said processing a category command comprises:

optionally adding a first one of the plurality of category nodes to a second one of the plurality of category nodes when the nodetype of the second one of the plurality of category nodes has a predefined type, the first one of the plurality of category nodes being a new category node, and the second one of the plurality of category nodes being an existing category node, said optionally adding a category further including;

accepting a user-defined category from an input device; storing the user-defined category within the new category node;

storing a nodetype within the new category node;

storing a unique NodeID corresponding to the user-defined category within the new category node;

storing a ParentID corresponding to a NodeID of the existing category node within the new category node; and

storing a LinkID within the new category node, the LinkID indicating a NodeID of one of the plurality of category nodes when the nodetype is of a predefined type.

4. The method according to claim 1, wherein said optionally processing a category command comprises:

optionally creating a link from a first one of the plurality of category nodes to a second one of the plurality of category nodes when the first one of the plurality of category nodes is not a root node and when it has a nodetype of a predefined type, the LinkID of the first one of the plurality of category nodes referring to the NodeID of the second one of the plurality of category nodes.

5. The method according to claim 1, wherein said optionally processing a category command comprises:

optionally moving one of the plurality of category nodes within the class hierarchy, said optionally moving further including altering the ParentID of the one of the plurality of category nodes.

6. The method according to claim 1, wherein said optionally processing a category command comprises:

optionally editing one of the plurality of category nodes within the class hierarchy, said optionally editing further including altering the category name of the one of the plurality of category nodes.

7. A computer system for creating a class hierarchy for categorization of documents within a memory, the class hierarchy for use with a document classification system capable of classifying a document based on content within the class hierarchy, the computer system comprising:

a processor; and

a memory having stored therein the following:

means for initializing the class hierarchy, the class hierarchy having a root category node within a tree data structure, the root category node having a user-defined category name;

means for displaying the class hierarchy;

means for accepting a user-selected command for manipulating the class hierarchy;

## 11

means for processing a category command in response to the user selected command having a first predefined state, causing the class hierarchy to contain a plurality of category nodes, said means for processing the category command further comprising: 5  
means for storing a category name in one of the plurality of category nodes, wherein each of the plurality of category nodes corresponds to a unique directory;

means for storing a NodeID within one of the plurality of category nodes, the NodeID defining the unique directory; 10

means for storing a nodetype within one of the plurality of category nodes, the nodetype when having a predefined type allowing a new category node to be added to a selected one of the plurality of category nodes, and otherwise preventing the new category node from being added to the selected one of the plurality of category nodes; 15

means for storing a ParentID within one of the plurality of category nodes, the ParentID indicating a NodeID of a parent category node; and 20

means for storing a LinkID within a first one of the plurality of category nodes, the LinkID indicating a NodeID of a second one of the plurality of category nodes when the nodetype is of a predefined type; and 25

means for processing a terms command in response to the user selected command having a second predefined state, the terms command manipulating terms defining one of the plurality of category nodes.

8. The computer system according to claim 7, the means for displaying further comprising:

means for displaying information corresponding to at least one of the plurality of category nodes within the class hierarchy, the information indicating a nodetype for the at least one of the plurality of category nodes. 30

9. The computer system according to claim 7, wherein the means for optionally processing a category command comprises:

means for optionally adding a first one of the plurality of category nodes to a second one of the plurality of category nodes when the nodetype of the second one of the plurality of category nodes has a predefined type, the first one of the plurality of category nodes being a new category node, and the second one of the plurality of category nodes being an existing category node, the means for optionally adding a category further including 40

means for accepting a user-defined category from an input device; 50

means for storing the user-defined category within the new category node;

means for storing a nodetype within the new category node; 55

means for storing a unique NodeID corresponding to the user-defined category within the new category node;

means for storing a ParentID corresponding to a NodeID of the existing category node within the new category node; and 60

means for storing a LinkID within the new category node, the LinkID indicating a NodeID of one of the plurality of category nodes when the nodetype is of a predefined type.

10. The computer system according to claim 7, wherein the means for optionally processing a category command comprises: 65

## 12

means for optionally creating a link from a first one of the plurality of category nodes to a second one of the plurality of category nodes when the first one of the plurality of category nodes is not a root node and when it has a nodetype of a predefined type, the LinkID of the first one of the plurality of category nodes referring to the NodeID of the second one of the plurality of category nodes.

11. The computer system according to claim 7, wherein the means for optionally processing a category command comprises:

means for optionally moving one of the plurality of category nodes within the class hierarchy, said optionally moving further including altering the ParentID of the one of the plurality of category nodes.

12. The computer system according to claim 7, wherein the means for optionally processing a category command comprises:

means for optionally editing one of the plurality of category nodes within the class hierarchy, said optionally editing further including altering the category name of the one of the plurality of category nodes.

13. A computer-readable medium recording software, the software disposed on a computer to perform a method for creating a class hierarchy for categorization of documents within a memory, the class hierarchy for use with a document classification system capable of classifying a document based on content within the class hierarchy, the method comprising:

initializing the class hierarchy, the class hierarchy having a root category node within a tree data structure, the root category node having a user-defined category name; 35

displaying the class hierarchy;

accepting a user-selected command for manipulating the class hierarchy;

processing a category command in response to the user-selected command having a first predefined state, causing the class hierarchy to contain a plurality of category nodes, said processing the category command further comprising:

storing a category name in one of the plurality of category nodes, wherein each of the plurality of category nodes corresponds to a unique directory; 45

storing a NodeID within one of the plurality of category nodes, the NodeID defining the unique directory;

storing a nodetype within one of the plurality of category nodes, the nodetype when having a predefined type allowing a new category node to be added to a selected one of the plurality of category nodes, and otherwise preventing the new category node from being added to the selected one of the plurality of category nodes; 50

storing a ParentID within one of the plurality of category nodes, the ParentID indicating a NodeID of a parent category node; and

storing a LinkID within a first one of the plurality of category nodes, the LinkID indicating a NodeID of a second one of the plurality of category nodes when the nodetype is of a predefined type; and 55

processing a terms command in response to the user-selected command having a second predefined state, the terms command manipulating terms defining one of the plurality of category nodes.

14. The computer-readable medium according to claim 13, said displaying further comprising:

## 13

displaying information corresponding to at least one of the plurality of category nodes within the class hierarchy, the information indicating a nodetype for the at least one of the plurality of category nodes.

15. The computer-readable medium according to claim 13, wherein said optionally processing a category command comprises:

optionally adding a first one of the plurality of category nodes to a second one of the plurality of category nodes when the nodetype of the second one of the plurality of category nodes has a predefined type, the first one of the plurality of category nodes being a new category node, and the second one of the plurality of category nodes being an existing category node, said optionally adding a category further including

accepting a user-defined category from an input device; storing the user-defined category within the new category node;

storing a nodetype within the new category node;

storing a unique NodeID corresponding to the user-defined category within the new category node;

storing a ParentID corresponding to a NodeID of the existing category node within the new category node; and

storing a LinkID within the new category node, the LinkID indicating a NodeID of one of the plurality of category nodes when the nodetype is of a predefined type.

16. The computer-readable medium according to claim 13, wherein said optionally processing a category command comprises:

optionally creating a link from a first one of the plurality of category nodes to a second one of the plurality of category nodes when the first one of the plurality of category nodes is not a root node and when it has a nodetype of a predefined type, the LinkID of the first one of the plurality of category nodes referring to the NodeID of the second one of the plurality of category nodes.

17. The computer-readable medium according to claim 13, wherein said optionally processing a category command comprises:

optionally moving one of the plurality of category nodes within the class hierarchy, said optionally moving further including altering the ParentID of the one of the plurality of category nodes.

18. The computer-readable medium according to claim 13, wherein said optionally processing a category command comprises:

optionally editing one of the plurality of category nodes within the class hierarchy, said optionally editing further including altering the category name of the one of the plurality of category nodes.

19. A computer data signal embodied in a carrier wave and representing sequences of instructions which, when executed by a processor, cause said processor to create a class hierarchy for categorization of documents within a memory, the class hierarchy for use with a document classification system capable of classifying a document based on content within the class hierarchy, by performing the following:

initializing the class hierarchy, the class hierarchy having a root category node within a tree data structure, the root category node having a user-defined category name;

## 14

displaying the class hierarchy;

accepting a user-selected command for manipulating the class hierarchy;

processing a category command in response to the user-selected command having a first predefined state, causing the class hierarchy to contain a plurality of category nodes, said processing the category command further comprising:

storing a category name in one of the plurality of category nodes, wherein each of the plurality of category nodes corresponds to a unique directory;

storing a NodeID within one of the plurality of category nodes, the NodeID defining the unique directory;

storing a nodetype within one of the plurality of category nodes, the nodetype when having a predefined type allowing a new category node to be added to a selected one of the plurality of category nodes, and otherwise preventing the new category node from being added to the selected one of the plurality of category nodes;

storing a ParentID within one of the plurality of category nodes, the ParentID indicating a NodeID of a parent category node; and

storing a LinkID within a first one of the plurality of category nodes, the LinkID indicating a NodeID of a second one of the plurality of category nodes when the nodetype is of a predefined type; and

processing a terms command in response to the user-selected command having a second predefined state, the terms command manipulating terms defining one of the plurality of category nodes.

20. The computer data signal according to claim 19, said displaying further comprising:

displaying information corresponding to at least one of the plurality of category nodes within the class hierarchy, the information indicating a nodetype for the at least one of the plurality of category nodes.

21. The computer data signal according to claim 19, wherein said optionally processing a category command comprises:

optionally adding a first one of the plurality of category nodes to a second one of the plurality of category nodes when the nodetype of the second one of the plurality of category nodes has a predefined type, the first one of the plurality of category nodes being a new category node, and the second one of the plurality of category nodes being an existing category node, said optionally adding a category further including

accepting a user-defined category from an input device; storing the user-defined category within the new category node;

storing a nodetype within the new category node;

storing a unique NodeID corresponding to the user-defined category within the new category node;

storing a ParentID corresponding to a NodeID of the existing category node within the new category node; and

storing a LinkID within the new category node, the LinkID indicating a NodeID of one of the plurality of category nodes when the nodetype is of a predefined type.

22. The computer data signal according to claim 19, wherein said optionally processing a category command comprises:

optionally creating a link from a first one of the plurality of category nodes to a second one of the plurality of

15

category nodes when the first one of the plurality of category nodes is not a root node and when it has a nodetype of a predefined type, the LinkID of the first one of the plurality of category nodes referring to the NodeID of the second one of the plurality of category nodes.

23. The computer data signal according to claim 19, wherein said optionally processing a category command comprises:

optionally moving one of the plurality of category nodes within the class hierarchy, said optionally moving fur-

16

ther including altering the ParentID of the one of the plurality of category nodes.

24. The computer data signal according to claim 19, wherein said optionally processing a category command comprises:

optionally editing one of the plurality of category nodes within the class hierarchy, said optionally editing further including altering the category name of the one of the plurality of category nodes.

\* \* \* \* \*

UNITED STATES PATENT AND TRADEMARK OFFICE  
**CERTIFICATE OF CORRECTION**

PATENT NO. : 6,055,540  
DATED : April 25, 2000  
INVENTOR(S) : William A. Snow and Joseph D. Mocker

Page 1 of 1

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

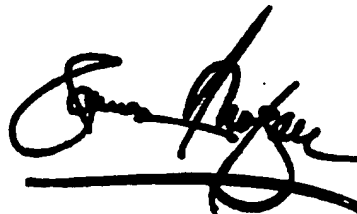
Column 3,

Line 20, replace "category node" with -- category/node --.

Signed and Sealed this

Sixteenth Day of April, 2002

Attest:

A handwritten signature in black ink, appearing to read "James E. Rogan", with a horizontal line drawn underneath it.

Attesting Officer

JAMES E. ROGAN  
Director of the United States Patent and Trademark Office